Data Science Methodology

A. Which topic did you choose to apply the data science methodology to? **(2 marks)**

Emails

B. Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. **(3 marks)**

You are required to:

1. Describe the problem, related to the topic you selected.
2. Phrase the problem as a question to be answered using data.

For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

1. Cybersecurity and privacy breach have become one the most common internet crimes over the past years. One of the most frequently encountered cyber crimes are phishing emails. Many people are not able to differentiate between an authentic email and a phishing email since the latter can very closely mimics the former. There is an increasing need to know the subtle differences between a real email and a phishing email.

2. How can we distinguish an authentic email from a phishing email by analyzing the anatomy and characteristics of an email ?

Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. **(5 marks)**:

1. Analytic Approach
2. Data Requirements
3. Data Collection
4. Data Understanding and Preparation
5. Modeling and Evaluation

1. We will take the predictive analytics approach, use the classification model and a decision tree to break down the components of an email and determine its authenticity.

2. We will require a huge sample size of both phishing and authentic emails. We will need to analyze the individual components of both of these two types of emails. In this particular case, the type of data we require will be mostly qualitative.

3. We will collect data from cyber crime prevention government and private agencies to get a sample size that is representative enough for the population of phishing and authentic emails. We will also need much data from colleges and universities because college faculty and students appear to be extremely susceptible to phishing emails and we can use these information to improve the predictive quality of our data science model.

4. We'll be primarily collecting the emails which are originally qualitative but we will later transform them to a more quantitative set of data we can better work with by running descriptive statistics and data visualization, which will allow us to be more familiar with the

data sets about the emails we have. Then, we will remove the data duplicates, sort the data based on certain criteria, fill the missing values and try to clean up the data so we can proceed to the modelling stage.

5. We will split the data about the phishing and authentic emails we have collected by splitting them between training and testing data sets in order to build a predictive model. We can then train the train data set using the train set and then predict the performance quality of our model using the test data set.

Finally, we can use the results from this data science model to build a simple web/mobile application or build an algorithm-based tool that can intelligently predict what the characteristics of a phishing email are.